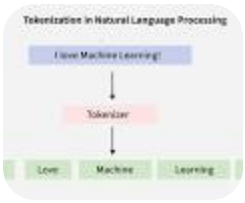
word tokenization

AI Mode All Images News Forums Short videos Web More Tools

AI Overview



Word tokenization is a foundational step in Natural Language Processing (NLP) that breaks raw text down into individual words or word-like units, known as tokens. This process transforms unstructured human text into discrete, manageable pieces that algorithms and large language models (LLMs) can process, analyze, and convert into numerical vectors.

For example, the sentence *"Machine learning is fascinating"* becomes the list of four word tokens: `["Machine", "learning", "is", "fascinating"]`.

How It Works

At its core, word tokenization relies on specific rules to determine where one word ends and the next begins. The process typically involves:

- **Delimiter Splitting:** Dividing text based on spaces, tabs, or newlines.
- **Punctuation Handling:** Stripping or isolating punctuation (e.g., periods, commas) so they don't attach themselves to words.
- **Case Normalization:** Converting all text to lowercase so the model treats "Apple" and "apple" identically, reducing vocabulary size.

Popular Tools and Libraries

You can easily perform word tokenization using standard Python libraries. Some of the most widely used tools include:

- NLTK (Natural Language Toolkit): Features the highly popular `word_tokenize()` method, which applies rule-based logic to English text.
- spaCy: An industrial-strength NLP library known for fast, robust, and language-aware tokenization.

- **Hugging Face Tokenizers:** Offers highly optimized tokenization pipelines essential for modern transformer models.

Common Challenges

While intuitive, word-based tokenization faces a few limitations that have led to the rise of *subword* tokenization (like Byte-Pair Encoding) in modern LLMs:

- **Contractions and Compound Words:** Handling contractions like "didn't" or compound phrases like "state-of-the-art" requires careful logic to avoid losing meaning or splitting them incorrectly.
- **Vocabulary Size and Out-of-Vocabulary (OOV) Tokens:** In highly inflected languages (like Turkish or German), there are limitless variations of words. Word tokenizers often struggle to understand new, unseen words, treating them as unknown.

Ask anything

I can provide **code snippets** to help you get started.

9 sites

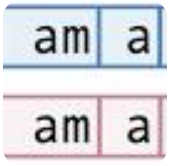
Tokenization in NLP - GeeksforGeeks

May 6, 2026 — Tokenization in NLP * Tokenization is a foundation step in NLP pipeline that shapes the entire workflow. * Involves dividing a str...

GeeksforGeeks

Words and Tokens - Stanford University

May 23, 2025 — Is the nature of words similar across languages? Some. languages, like Vietnamese or Cantonese, have very short words while others...



Stanford University

✅ Beginner's guide to Word Tokenization 📑 ✅ - Kaggle

What is word tokenization? ¶ Tokenization is breaking the raw text into small chunks. Tokenization breaks the raw text into words,

Kaggle

Show all

Stanford University
https://web.stanford.edu › ~jurafsky › slp3 › 2.pdf

Words and Tokens

PDF

May 23, 2025 — **tokenization**, the task of separating out or **tokenization** tokenizing **words** and **word** parts from running text. the running input text into tokens.
36 pages

GeeksforGeeks
https://www.geeksforgeeks.org › nlp › nlp-how-tokeniz...

Tokenization in NLP

May 6, 2026 — Word tokenization is the most commonly used method where **text is divided into individual words**. It works well for languages with clear word ...

People also ask

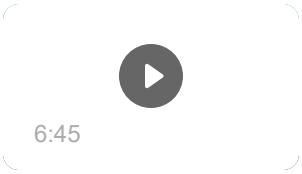
What is an example of a word token?

How many words are 1000 tokens?

What is tokenization of words in AI?

What is an example of word_tokenize?

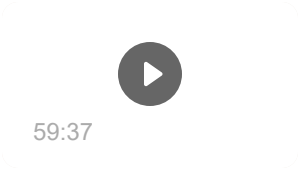
Videos



4 NLP: How to convert text into Tokens (Tokenization) ?

YouTube · SachinDataEngineer
Feb 28, 2025

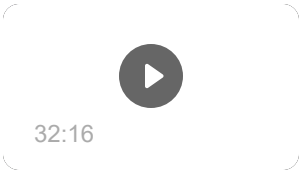
4 key moments in this video



Tokenization in NLP: Basics to Advanced Techniques ...

YouTube · Data Science Dojo
Jun 27, 2024

10 key moments in this video



310 - Understanding sub word tokenization used for NLP

YouTube · DigitalSreeni
May 3, 2023

10 key moments in this video

View all

DataCamp
https://www.datacamp.com › Blog › Artificial Intelligence

Tokenization in NLP: How It Works, Challenges, and Use ...

Word tokenization. This method **breaks text down into individual words**. It's the most common approach languages with clear word boundaries like English.

What Is Tokenization? | Types of Tokenization | Languages without clear...

Towards Data Science
https://towardsdatascience.com › Latest

Word, Subword, and Character-Based Tokenization

Jul 1, 2021 — **Tokenization** in simple **words** is the process of splitting a phrase, sentence, paragraph, one or multiple text documents into smaller units. Read more

IBM
https://www.ibm.com › docs › watson-libraries › topic=...

Tokenization and Word Segmentation

As a general principle, **Watson NLP follows the tokenization of UD corpora**. However tokenization and word segmentation are not trivial tasks in many ...

arXiv
https://arxiv.org › cs

Word-Pooled Tokenization for Language Modeling

by A Thawani · 2023 · Cited by 16 — This paper considers an alternative 'learn your tokens' scheme which utilizes the word boundary to **pool bytes/characters into word representations**. Read more

McKinsey & Company
https://www.mckinsey.com › mckinsey-explainers › wh...

What is tokenization?

Jul 25, 2024 — **Word tokenization splits text into individual words or word-like units**, and each word becomes a separate token. Word tokenization might struggle ... Read more

Supervised Machine Learning for Text Analysis in R
https://smltar.com › tokenization

Chapter 2 Tokenization

Word-level tokenization is done by finding word boundaries according to the specification from the International Components for Unicode (ICU). How does this ... Read more

People also search for

Sub word tokenization	**Types of** tokenization **in NLP**
Tokenization **NLP example**	**Subword** tokenization **example**
Word tokenization **nltk**	Tokenization **code**
Word tokenization **example**	Tokenization **algorithm**

Canada **Ville-Marie, Montreal, QC** - From your IP address - Update location

Help Send feedback Privacy Terms